

Historic, archived document

Do not assume content reflects current scientific knowledge, policies, or practices.

G-217.9

STATE-INITIATED FNP DEMONSTRATION PROJECT
ASSISTANCE AND EVALUATION

VOLUME III: FINAL REPORT

August 31, 1988

Prepared for:

Office of Analysis and Evaluation
Food and Nutrition Service
U.S. Department of Agriculture

Prepared by:

Applied Management Sciences, Inc.
962 Wayne Avenue, Suite 701
Silver Spring, Maryland 20910

TABLE OF CONTENTS

<u>Chapter</u>		<u>Page</u>
1	INTRODUCTION	1.1
2	MARYLAND PROJECT	2.1
	2.1 Overview	2.1
	2.2 Design of the Intervention	2.2
	2.3 Research Design	2.4
	2.4 Implementation of the Treatment	2.10
	2.5 Implementation of the Research Design	2.18
	2.6 Results and Conclusions	2.21
3	NORTH CAROLINA PROJECT	3.1
	3.1 Overview	3.1
	3.2 Design of the Intervention	3.2
	3.3 Research Design	3.6
	3.4 Implementation of the Treatment	3.12
	3.5 Implementation of the Research Design	3.16
	3.6 Results and Conclusions	3.19
4	VERMONT PROJECT	4.1
	4.1 Overview	4.1
	4.2 Design of the Intervention	4.3
	4.3 Research Design	4.7
	4.4 Implementation of the Interventions	4.9
	4.5 Implementation of the Research Design	4.11
	4.6 Results and Conclusions	4.12
5	COMPARATIVE ASSESSMENT	5.1

INTRODUCTION

As the responsible agency for administering the U.S. Department of Agriculture's Food Stamp Program, a major priority for the Food and Nutrition Service (FNS) is to reduce losses to the program due to fraud and error. It has become increasingly difficult to develop strategies that are effective in reducing errors. As the error rate decreases nationally, there remains a very costly set of errors that represent a relatively small percentage of total cases. Further, these errors are distributed heterogeneously, and they are elusive to methods that have been effective in the past. FNS has sponsored several initiatives to reduce errors. One of the most significant resources for error reduction strategies is state and local food stamp agencies. By virtue of having responsibility for service delivery, state and local agencies have firsthand knowledge of the systems that allow errors to occur. This knowledge cannot be duplicated within FNS or by external sources.

In July of 1983, USDA solicited state and local agencies to submit proposals for demonstrations and evaluations designed to reduce errors and abuse in the Food Stamp Program through fraud prevention and/or detection strategies and improved management practices. Authorization for this solicitation is contained in Section 17 of the Food Stamp Act, as amended. FNS established cooperative agreements with three states--Maryland, North Carolina, and Vermont--as a result of this solicitation. The Maryland project proposed a media presentation that was designed to prevent errors by informing applicants of reporting requirements and encouraging them to report accurately. North Carolina hypothesized that errors occurred because caseworkers failed to be comprehensive in taking application information and

in noting inconsistencies in the information provided. To remedy this problem, North Carolina proposed a computer-assisted interview which would require a response to each application question, and would compare responses to assure that consistent information was reported. Vermont noted that a substantial proportion of its errors were due to agency deficiencies and developed four strategies for reducing error: a supervisory case review system, enhanced eligibility worker training, quality circles, and the establishment of performance objectives.

The purpose of this report is to summarize and assess these three demonstrations. The chapters that follow cover the following topics for each demonstration:

- Overview
- Design of the Intervention
- Research Design
- Implementation of the Treatment
- Implementation of the Research Design
- Results and Conclusions.

This is followed with a chapter that comprises the demonstrations and comments on how their differences may have affected outcomes.

MARYLAND PROJECT

2.1. OVERVIEW

The Maryland demonstration was intended to reduce client error by exposing clients to media presentations--a videotape and a brochure--in local office waiting rooms. This preventive strategy is based on the assumption that clients withhold or falsify information because of ungrounded fears that not doing so will lead to ineligibility. Accordingly, the media messages were benign, rather than threatening.

The demonstration team developed and implemented a sound research design, although the sample size was probably inadequate. They also developed, through a subcontract, satisfactory media presentations. Despite extraordinary delays, the demonstration itself was well conducted and led to a sample size equal to about 75 percent of the target value.

The subsequent activities of data collection, editing, coding, and key entry were problematic. This is partly attributable to project staff who, although conscientious and diligent, were not adequately prepared for dealing with real-world data. A more salient factor was the isolation of these staff from State and local employees who might have provided help and/or leadership; indeed, the office conducting the demonstration appeared to be isolated from other units in the State Food Stamp Program. Finally, although we are not in a position to assess Maryland relative to other States, it appears that the general flow of data and the linkages among agencies in this State are not conducive to research activities.

Delays in data collection, and then in preparation of a computer file, left less than two weeks for analysis and interpretation. As a result, these activities were not done adequately. In addition, what was done was not reported adequately. As a result, there is no analytic basis for assessing the effect of the intervention on the outcome variables, let alone on the QC error rate.

2.2 DESIGN OF THE INTERVENTION

Specification of the Problem

Maryland utilized QC data to decide that client error was to be the broad target of the demonstration and that, more specifically, client reporting of income--both earned and unearned--and of household size and composition were the major error sources. Data were examined for several recent QC cycles, and it seemed clear that there was a consistent pattern.

Although some of this examination took place prior to funding of the project, the greater share of it occurred after the project was underway. It would have been preferable for Maryland to analyze error sources more thoroughly at the time that they submitted their proposal. Nevertheless, the analysis was rational and useful, and the decision to target client error was a good one.

No attempt was made to identify error-prone client types. One reason for this was that the office that initiated the project appears to have no communication with the QC staff, so that it would have been difficult to obtain QC microdata or even to request QC staff to supply more detailed reports than are normally available. (More generally, there is an apparent lack of communication between the initiating office and several divisions and levels. A more unified approach to developing an intervention would have been helpful. This issue will reappear later in the discussion.)

A second reason for not targetting client types was that the decision was made early on to develop an intervention that, by its very nature, applied to all clients and applicants. Thus, it did not appear relevant to focus on

particular types. Nevertheless, it would probably have been beneficial to investigate this issue more closely. It might have led to more insightful media messages.

Definition of the Intervention

There are several ways in which client errors can arise. At one end of the spectrum, applicants may simply be guilty of willful misreporting. At the other end, they may not be aware of reporting requirements. It seems likely that, at least for many clients, the truth lies somewhere in between these extremes. This was the assumption that Maryland made. The State hypothesized that applicants were not consciously trying to cheat the system, but were misreporting because of fear that more complete and accurate reporting would reduce grants or render them ineligible.

Given this hypothesis, the most obvious corrective strategy is that of allaying client fears by informing clients that accurate reporting will not have negative consequences. Maryland took this approach, and designed a preventive treatment rather than a corrective one. The treatment was conceptualized and developed in advertising terms, as media messages intended to change clients' mindsets so that they would be more likely to report accurately. The messages were designed to reinforce the existing reporting requirements, and to make it clear that complete reporting did not necessarily have negative consequences.

This approach is internally consistent and is humanistic. However, it reflects the obvious fact that reporting more may well lead to receiving less. Thus, it was not at all clear that the media messages would have the desired impact. In fact, it seems possible that they could have an impact in the wrong direction: By coaxing clients to report more, Maryland may be reinforcing the implicit message that the State is frequently unable to detect incomplete reporting. There is no firm basis for arguing on one side or the other of this issue. However, it should be noted that project staff in Maryland were never very optimistic about the effects of the treatment, and that local office staff--who are probably best qualified to predict the effects--were not consulted until the project was underway. It would have been better to develop an intervention with more face validity.

The intervention was originally intended to include a video presentation and two brochures. Various combinations of these were to be implemented. During initial planning, this was reduced to a video presentation and one brochure, with the actual treatments being the brochure only and the brochure in conjunction with the video. This reduction in scope was warranted for two reasons. First, the sample size was inadequate for testing many treatments and, second, there was no strong conceptual basis for fine-tuning the treatments to the extent originally proposed.

Estimation of the Potential Benefit

There was no basis for estimating the reduction in error that might be expected as a result of the media treatments. Since the direct effect could not be estimated, neither could the ultimate effect on the QC error rate. This is a generic problem that applies to all or almost all prevention strategies, and no specific criticism of the Maryland project is implied here. Nevertheless, as mentioned above, it would be better to fund projects that, in the absence of valid numeric estimates, are at least believed to offer more potential for error reduction.

With no basis for estimating the treatment efforts, there is no basis for estimating cost effectiveness. However, it should be noted that projects of this type, if effective, are bound to be cost effective. The reason for this is that, because the statistical power for detecting an effect is relatively small, any effect that is detected must be relatively large. Since the ongoing costs of the treatments would be very small, the benefits would almost certainly outweigh the costs.

2.3 RESEARCH DESIGN

Definition of a Control Group

The first impulse, when looking for a way to measure the treatment effect of an error reduction project, is to rely on QC error rates. The impulse is a natural one, since the ultimate goal is to reduce this rate. It should be avoided, however, as past QC samples are not adequate control groups.

Comparisons between past and present QC rates are unreliable because the samples are too small, and are generally invalid because of intervening events other than the demonstration. Thus, although it is relevant to track QC results, it is also necessary to develop a design that is specifically linked to the demonstration.

The Maryland Project recognized this necessity, and proposed control groups at each of the six local offices--two each in Baltimore City, Prince George's County, and Montgomery County--at which the experimental treatments were to be implemented. Each group was to be selected during a particular block of time, with the control group selected first, then the brochure-only group, and then the brochure-plus-video group. Within a given block of time, there was no randomization, but rather an attempt to oversample and include as many subjects as possible.

This design, while not a true experiment, appears to be satisfactory. The ordering of the treatments avoids contamination. Other threats to validity are possible, but not likely. These include:

- Changes in the client mix from one block of time to the next;
- Changes in local office staff which could influence client reporting;
- Operational or policy changes.

The second and third of these possible threats apparently did not arise. The first seemed to be a possibility, particularly with regard to the mix of applicants and recertification cases, and indeed it turned out that this mix did differ in a consistent way across the groups.

With regard to client mix, it was suggested to the Maryland Project that the samples be balanced, either across offices or at each office separately, on variables that might interact with the treatments. These include the applicant/recertification and public-assistance/non-public-assistance variables. The suggestion was not followed, because it was believed that it would be difficult for local office staff to properly implement it, and because there was no clear reason to believe that interactions would occur.

While this should not be considered a major flaw, it would certainly have been preferable to involve local office staff at the planning stage, and to assure a more balanced sample.

Development of Outcome Measures

The basic variable of interest to the Maryland Project was overpayment due to client misreporting. As mentioned earlier, there was evidence that the two major areas in which clients were misreporting were income and household composition. In addition, vehicle ownership was included as a third area. Finally, income was separated for operational purposes into earned and unearned income. Thus, there were four areas in which misreporting was to be studied.

The ideal approach in this situation would be to:

- (1) Use verification procedures to determine true household composition, true earned income, etc.
- (2) Compare verified data with reported data and note discrepancies
- (3) When discrepancies occur, recompute the grant amount and note the amount of overpayment.

This ideal was not pursued in Maryland for several reasons. First, project funds did not allow for recomputation of grant amounts, and, as far as we know, local office staff were not interested in following up in this way. Second, obtaining valid dollar figures through verification turned out to be difficult in some areas, and impossible in others. The key consideration here was determination of earned income. The project had originally intended to measure this, but later decided that it was not possible to do so. Third, there was no feasible procedure for determining true household composition, except in the case of school age children.

These problems led to two decisions. First, rather than attempting to simulate QC procedures, the research design was predicated upon the use of indicators, each indicator being thought of as evidence of a possible discrepancy between reported and true data. Second, in addition to

discrepancy indicators, client reported data per se were to be used as outcome variables. In other words, extent of reporting was included as a second generic variable, based on the assumption that the more information that is volunteered, the more accurate the total information is likely to be.

The strategy of collecting data for indicators and surrogate measures is acceptable. Indeed, it seems to be the only feasible strategy that Maryland could have used. It is unfortunate, however, that this was not realized at the outset: too much time was spent in developing a measurement scheme. Better linkages between project staff and QC staff, and between project staff and local office staff, would have alleviated this problem.

The key outcome measures that were finally selected include:

- Dichotomous measures of discrepancies between reported and verified data in each of the four content areas
- Dichotomous measures indicating presence or absence of earned income, unearned income, and motor vehicles.

These measures are either defined at the household level or defined at the individual level and recorded for the household. As a final step, discrepancy measures can be combined across areas to yield a single variable for a household.

Sample Sizes

Sample sizes should be large enough to enable detection of a treatment effect of reasonable size. To operationalize this concept, it is necessary to define the treatment effects and their expected sizes. Treatment effects, in turn, follow from research hypotheses, which specify comparisons of outcome measures between groups.

Maryland developed adequate research hypotheses of two types. First, there were straightforward comparisons, i.e., comparisons between two groups on individual variables. Second, there were more complex hypotheses,

involving the partialing out of background variables and other multivariable techniques. Hypotheses of the first type can, and were; used to analyze the issue of sample size.

Unfortunately, the total sample size was set at 1,200 prior to the analysis. Therefore, the issue became one of justifying (or not justifying) a preset number, rather than one of rationally determining a number.

As mentioned earlier, one reason for reducing the number of experimental treatments was the fixed total sample size. As the number of treatments decreases, the number of cases per treatment increases, and the power of any particular comparison increases correspondingly. For two treatments, together with a control group, there could be 400 cases per group. Alternatively, it might have been preferable to assign half of the sample to the control group and half to the experimental groups--300 to the brochure-only and 300 to the brochure-plus--to maximize the power of the overall tests of experimental versus controls. Maryland opted for the former assignment, because of cost considerations: Because data collection is more costly than exposure to the treatments, it is preferable to minimize the size of the control group.

Suppose, to gauge the adequacy of the sample sizes, that the client error rate (on a case basis) is initially 10 percent and that the treatments, taken together, can be expected to reduce this rate by 25 percent of its initial value. Then a straightforward test for a difference in proportions could be expected to yield proportions of 10 percent in the control group and 7.5 percent in the experimental group. With sample sizes of 400 and 800 in the two groups, the difference would have a standard error of:

$$\text{SQRT}\left(\frac{.1 \times .9}{400} + \frac{.075 \times .925}{800}\right) = .0177$$

The corresponding Z-score is .025/.0177, or 1.42, which is not statistically significant at the 95 percent confidence level. It follows that the probability of a Type II error--failing to detect a treatment effect when it does occur--is greater than 50 percent. Thus, the experimental setup has inadequate power.

This argument rests on a number of assumptions, the most salient of which is the assumed reduction of 25 percent in the client error rate. If a higher amount of reduction were assumed, the power would increase. There is no evidence, however, to suggest that even a relative reduction of 25 percent could reasonably be expected. We conclude, therefore, that it was unlikely at the outset that the demonstration would successfully show evidence of error reduction.

It is important to note that the problem here was not the statistical one of failing to do an a priori power analysis, and thus having a sample that was probably inadequate. The problem was the more basic one, discussed earlier, of face validity. The post hoc power analysis simply focuses attention on the relatively large reduction in error that corresponds to the given sample size.

The above remarks apply to the most global group comparisons: those for which the entire experimental group is compared to the entire control group. In fact, the project was interested in many lower-level comparisons, all of which involved smaller samples. The most important of these were based on keeping the two experimental groups distinct. Comparisons of this type involved 400 cases per group, leading to a reduction in expected Z-score from 1.42 to 1.25.

Other comparisons that were of interest involved applicants only, recertification cases only, public assistance cases only, etc. The typical cut of this type would lead to group sizes of 200 controls and 400 (combined) experimental, and a resultant Z-score of 1.00. Cuts based on two or more variables would produce expected Z-scores of less than 1.00.

The proposed multivariate analysis (log linear analysis) could resolve problems of differential treatment effects by controlling for various background variables. It could not, however, resolve the basic issue of detecting a treatment effect that applies only to specific subgroups.

Generalizability

The original intent of the Maryland project was to implement the demonstration in a "representative" sample of counties. This was

operationalized by stratifying counties into "large", "medium" and "small," the plan being to select one or two counties from each stratum. (Note that Baltimore City qualifies as a "county" for this purpose.) There is no methodological fault to be found with this plan.

The project subsequently decided to limit the demonstration to three large counties: Baltimore City, Montgomery County and Prince George's County. The State was motivated in this decision by the fact that the preponderance of errors--over 80 percent, in fact--occur in these three counties. Thus, the demonstration, if it were successful, would have a relatively large impact on the error rate. This was a sound decision from an operational standpoint, but not necessarily from a research standpoint, since it sacrificed both randomness and representativeness. The results, in other words, could not be validly generalized to smaller counties, to the State of Maryland, or to other States.

In fact, however, the decision was a wise one, methodologically as well as operationally. There are two reasons for this. First, going into the larger local offices minimizes the risk of a shortfall on the sample sizes. Second, it maximizes the likelihood of detecting a treatment effect. Since these are the offices with the higher error rates, they offer more "room for improvement": it is easier to detect a change in error rate from 10 percent to 7.5 percent, as opposed to a change from, say, 5 percent to 3.75 percent.

Furthermore, the loss in generalizability seems minor. If the demonstration were successful at the selected sites, it could be expected to be successful at similar sites, whether in Maryland or in other States. Finding strategies that work in urban areas, where error rates are typically higher to begin with, is probably more important than testing strategies on a representative basis across counties.

2.4 IMPLEMENTATION OF THE TREATMENT

Obtaining of Cooperation

The Maryland Project obviously required a high degree of cooperation from local offices. In addition, it would have been very beneficial, although not

an absolute necessity, to have had a working relationship with other offices at the State level: The Quality Control Staff, the AIMS Staff, and the Procurement Staff. The issues involving local offices will be discussed first.

Ideally, local offices should not merely cooperate with a State initiative, but should have some input into the development of that initiative. This did not happen during the initial planning stages in Maryland: The use of a brochure and a videotape, and the general content of the brochure and videotape, were based entirely on State decisions. Thus, the counties had no initial sense of "ownership" of the demonstration.

Local staff were, however, consulted fairly early in the project to ensure that the intervention was at least feasible from an operational standpoint. And, they were consulted during the development of the videotape to get their opinions on the content and the tone of the message. Thus, a working relationship and some sense of partnership were fostered.

In operational terms, local offices were requested to do the following:

- (1) Maintain log-in sheets for each of the three groups: Control, brochure-only, and brochure-plus video
- (2) Use the log-in sheets to draw the three samples, and forward copies of the corresponding applications to the project
- (3) Distribute brochures to the two experimental groups
- (4) Maintain the video presentation and advise clients (from the brochure-plus-video group) of its availability.

These responsibilities do not seem very burdensome, particularly since the demonstration period for each group at each local office was only one week. Nevertheless, the required activities were outside the normal sphere of operations, a succession of minor but annoying problems could be expected to arise, and the project offered no "payoff" to the staff who would actually be doing the additional work. Therefore, complications were to be expected.

To pave the way toward a smooth flow of events at the local offices, project staff did a very conscientious job of explaining what would be

required. These explanations were presented to supervisory staff at each office, and supervisors expressed willingness to cooperate. Thus, all reasonable steps were taken.

Two additional points need to be stressed here. First, the Maryland Food Stamp Program is county-administered. Therefore, cooperation with regard to a State demonstration is a very real issue: County cooperation must be solicited rather than mandated. Second, the State office that conducted the project appeared to have virtually no linkages to either the county structure or the operational offices at the State level. Furthermore, almost all contact with local offices went through the project staff who were contract employees, rather than the project director. In light of these impediments, the project staff did an exemplary job of obtaining cooperation from the local offices.

As mentioned at the beginning of this section, it would have been desirable for the project to have had a working relationship with other offices at the State level. Most importantly, since the research design required verification sources external to the office conducting the demonstration, there should have been assurances that these sources would "work." Furthermore, when computer screens from the AIMS (Maryland's automatic food stamp case processing system) became available, and seemed to constitute another verification tool, a working relationship with computerization staff should have been developed. This too did not happen, although some ad hoc help was received from people familiar with AIMS.

Finally, since procurement of services from a media subcontractor was a vital component of the project, there should have been assurances that these services could be obtained in a timely way. Instead, all steps of the process--development of an RFP, evaluation of proposals, and finalization of the eventual subcontract--were burdened by lengthy delays and confusion over administrative procedures. It seemed for a while that these delays might lead to termination of the whole project. Similar, but not quite as serious, problems arose with regard to purchasing software and renting one micro-computer for two months. Admittedly, procurement can be a frustrating process in a government bureaucracy. Nevertheless, the lack of cooperation that

emerged seemed excessive. This appeared due, not to any shortcomings of the project staff, but to a general isolation of the State office conducting the demonstration from the main infrastructure of the State agency.

Pilot Testing

The Maryland Project conducted a pilot test of the data collection procedures after obtaining cooperation from the six local offices, but before developing the brochure and videotape. The purposes of the pilot test were to determine the feasibility of obtaining data from the local offices and from the verification sources, and to see if any modifications were necessary.

Each of the local offices were asked to supply photocopies of ten applications: Five from applicants and five from recertification cases. The resultant data were coded and used to generate verification requests from the verification sources: Employment Security, Social Security, The Motor Vehicle Bureau, and local school districts.

A number of problems emerged. First, the application forms were slow to arrive. There was no specific reason for this, and it indicated that, despite assurances from supervisors, the workers who actually pulled the files were busy people with little time to spare for special projects. Second, it turned out that different application forms were in use: The standard food stamp application was in general use, but Baltimore City used a special form for recertification, and other counties apparently used the Public Assistance (AFDC) Form as well as the Food Stamp Form. This proliferation of forms was unexpected and led to problems in coding.

The third problem was one of sparse data. Applications tended to have little substantive information, and, not infrequently, had missing or inconsistent identifying information. The paucity of substantive information suggested that the ultimate data analysis might be tenuous: It is difficult to show differences between discrepancy rates when very few clients have anything to report. The problems with names and social security numbers suggested that verification activities would be problematic: There would be difficulties in matching applicants to verification files.

The final problem was with the verification sources themselves. These sources appeared to be unreliable and incomplete. The motor vehicle file was particularly troublesome: There were more clients who reported vehicles not on the file than there were clients with unreported vehicles that were on the file.

Project staff devoted a great deal of time to resolving these problems. They learned how to interpret ambiguities on the application forms, how to establish linkages between different forms, how to request verification data, and how to interpret these data. They rethought the whole concept of a "discrepancy," and developed practical rules for identifying and coding discrepancies. And, they were able to formulate a more realistic notion of what might be expected from local office staff.

In summary, the pilot test was very useful. It did not lead to any one major change in plans, but it did produce many minor modifications, and lead to a more realistic set of expectations.

Project Scheduling

The Maryland Project was delayed at its inception, and continued to experience delays throughout its duration. The initial problem was one of staffing: More than six months were required to hire a Project Manager, and several additional months went by before a second professional was hired.

Subsequent delays can be attributed to numerous causes. The procurement problem, discussed earlier, was certainly a major cause. Other contributing factors include:

- An inordinately large amount of time to develop a mutually acceptable research and analysis plan
- Unexpected complications with regard to the verification process
- Unanticipated delays in receiving data from local offices and from verification sources
- A complicated--possibly over-complicated--system for coding and transcribing data

- Inadequate access to computers.

It should not be inferred that the State bears sole responsibility for the delays. FNS, as well as its technical assistance contractor, could have imposed more intervention to maintain the project schedule and to anticipate problems before they became serious. However, it was decided that the roles of FNS and the technical assistance contractor would be to provide advice and guidance, but leave it to the Project Director to carry out the advice and guidance.

It is important to note that, although the timelines were considerably extended, the demonstration was fully implemented. The delays did not lead to any "shortcuts" or curtailing of planned activities.

Implementation Per Se

Implementation consisted of two major components: Developing a brochure and a videotape, and using these media at the six participating local offices. Once the media subcontract was signed, development went fairly smoothly. Preliminary versions of the videotape were critiqued by project staff, by FNS staff, and by a panel of local workers and clients. The final version incorporated suggestions made by these groups.

The message of the videotape is: Be honest, report accurately and completely, and you will get the benefits that you deserve, which may be more than you would have guessed. This may not be a very persuasive message. However, it is the one that the project wanted to make, and it was implemented in a competent and professional manner by the subcontractor.

The brochure is more factual. It provides specific guidelines for what the client is obligated to report. Again, it was designed and produced competently and professionally.

Clients were exposed to each experimental treatment--brochure-only and brochure-plus-video--for one week at each of the six participating local offices. The brochure-only week came after the control-group week and before the brochure-plus-video week to avoid contamination effects.

The treatments were staggered across offices so as to permit one demonstration staff member to be at a local office for most of the time that the demonstration was in progress. This person monitored the demonstration-related activities in the client waiting area: Use of the log-in sheet, distribution of the brochure, and maintenance of the video. The role was a passive one: Staff members did not interact with clients, but only with receptionists, and, when necessary, with supervisors.

There were some problems with using the log-in sheets and distributing the brochures. Although cooperation had been promised at each office, the message did not necessarily reach the person actually doing the work: The receptionist. Some receptionists did not know what they were supposed to do, and others considered the additional work to be somewhat of an imposition. Problems at this level were generally resolved quickly, and logging in and brochure distribution were adequately and uniformly implemented. It should be stressed that monitoring by project staff was essential; without this monitoring, implementation would have been seriously flawed.

The video presentation raised additional problems. When waiting rooms were noisy, it seemed desirable to increase the volume of the presentation so that the clients who wanted to could hear it. When the ambient noise level dropped, the volume seemed too high, and clients were likely to turn it down--or, in some cases, to turn it off altogether. The problems were exacerbated by the layout of the waiting rooms. In some offices, decreasing the volume to accommodate clients sitting near the video player resulted in clients further away not being able to hear.

Another factor to consider is that, in more crowded offices and on busy days, clients are frequently in the waiting area for over an hour. The video presentation can become quite intrusive, even for someone who is initially receptive, after six or seven viewings. And, the intrusion is obviously more pronounced for a receptionist who is stationed within listening range.

Project staff monitored the video closely, and tried to maintain the volume at a reasonable level. This monitoring was clearly necessary, but could not resolve all problems. Although the video component was implemented

as fully as possible, it could not be called uniform across sites. The client mix, the density of clients, the attitude of the receptionist, and the physical layout of the waiting room all contributed to variability in the exposure of the presentation to the clients.

The question of exposure is an important one in demonstrations of this type. If a media message is to change a client's behavior, the message must first reach the client. The Maryland Project, realizing this, initially wanted to have a project staff member handing out brochures and directing clients to the video presentation. This would have increased the proportion of clients who were exposed, and presumably would have augmented the treatment effects. It would not, however, have been generalizable to a non-demonstration setting, and hence the idea was discarded.

It was then suggested that an exposure scale be developed, and that observers rate clients on this scale. This might have permitted a partialing out of the exposure variable, and led to estimates of "true" treatment effects. It was decided, however, that exposure was too difficult to measure, at least within the confines of this project. Furthermore, this "true" effect is of only academic interest. The practical concern is not the effect of watching a video, but the effect of having had the opportunity to watch a video.

Although there was no formal attempt to measure exposure, it is certainly of interest to summarize clients' reactions to the media. With regard to the brochure, most clients accepted it as an addition to whatever they were carrying. Some looked through it; most did not. There is no way of telling the extent to which they subsequently read it.

With regard to the video, many clients looked at it sporadically. Some seemed to be watching it intently; others seemed to ignore it. Again, there is no way to determine the extent to which they received the message.

2.5 IMPLEMENTATION OF THE RESEARCH DESIGN

Data Collection

In collecting data for subsequent analysis, Maryland confronted a number of problems. These can be divided into two categories: problems related to the reported data, i.e., the data recorded on the application form, and problems related to the verification processes that were undertaken by project staff.

Data on the application forms appeared to be of low quality: omissions, illegible identifying information, and other similar problems were common. We do not have sufficient experience with other States--at least, not at the level of detailed analysis of application forms--to put this in a proper (national) perspective. Thus, we cannot discount the possibility that food stamp applications in Maryland are about on a par with those in other States, with regard to data quality. Perhaps it is more to the point to reiterate that project staff were themselves unfamiliar with the realities of application data, and that the linkages between the project and the food stamp infrastructure--both local and State--were tenuous. This may have led to an over-reaction on the part of project staff. At any rate, they spent much time and energy on transcribing data, developing rules for handling a great variety of contingencies, recoding data, and generally attempting to portray the dataset as viably as they could. Unfortunately, we cannot determine the net result of this effort at the present time: the final report that Maryland has delivered does not address the issue of data quality.

One tactic that project staff adopted was to use AIMS screens to validate application data, and to expand these data. Again, this was a time-consuming effort, and it is not at all clear that it was worthwhile.

A particular issue to note in the present context is that of whether, for public assistance (PA) cases, the food stamp application is even supposed to be complete. During the early stages of the project there was a question of whether the food stamp application was intended as a stand-alone document, or

was, for PA cases, an adjunct to the AFDC application. We had believed this issue to be resolved, and were told that the food stamp application is indeed complete, whether or not it is associated with an AFDC application. The final report delivered by the project, however, States that this is not the case, and goes on to mention that, because of incomplete food stamp applications for PA clients, various blank fields have been coded as missing, rather than as zero. This can be expected to influence the results of the data analysis, but there is no information provided that could enable us to investigate the extent or consequences of the problem.

A final issue with regard to client reported data relates to the use of different forms for different clients. Specifically, there is a special form used for all recertification cases in Baltimore city. The "BC Recon Form," as it is commonly called, is similar in broad terms to the standard application--the FSI--but differs somewhat at the level of individual data elements. This did not seem to be a serious problem to us, but was considered serious by project staff. Eventually, they created separate coding forms, data entry programs, etc., for the two classes of sampled clients, a procedure that was very time-consuming and possibly unnecessary. At any rate, the two sources of reported data were reconciled: the analytic data file contains the same set of elements for all subjects.

With regard to the verification process, the general problem is addressed in the final report prepared by Maryland: verification data are of questionable quality because of problems in matching identification fields, the time lag between verification sources and reported data, and the failure of verification information to adequately address the reporting dimensions that were of primary interest.

The verification problems were obvious from the pilot test, which indicated that discrepancies were more likely to occur in the "wrong" direction than in the "right" direction. The result of the motor vehicle registration verification, for instance, was that several reported cars did not appear on the verification file, but no unreported cars were detected.

In summary, the data collected by the project are not highly reliable, and discrepancy measures constructed from these data are somewhat suspect. Measures that capture the extent of reporting are more valid than discrepancy measures, and should probably be the focus of any additional analysis conducted by FSN.

Data Analysis

Data analysis generally followed the steps that had been presented in the research plan. However, because there was very little time for conducting analysis, there was no opportunity for project staff to review and interpret results, and to modify subsequent analyses accordingly. Analyses were performed on a personal computer, using SPSS, and included contingency table analysis, analysis of variance, and log linear analysis. As far as can be determined, virtually no basic descriptive analysis was done, nor was there any attempt to calculate standard errors or to introduce costs into the analysis.

Reporting

The final report submitted by the State of Maryland is inadequate. It is largely comprised of previously submitted material and does not seriously address the issue of whether or not the intervention resulted in lower error rates. The report States that the research was inconclusive. This may well be the case, but all that can be said at the present time is that the material presented in the report is inconclusive.

There is some discussion of problems that were encountered during the demonstration, and of factors that adversely affect the quality of the data. However, there is no attempt to measure the impact of these factors, a task that would be fairly straightforward given the data quality variables that were developed and presumably are included in the analytic dataset.

The presentation of analytic findings is particularly disturbing. Many significant effects were obtained from analyses of variance and log linear analyses, but the summary tables that are presented do not permit the reader

to determine the directionality of the differences. Most of these significant differences are dismissed in the accompanying text as statistically significant but too small to be of practical concern. This may or may not be the case, but the reader should be permitted to judge for herself.

Maryland has provided FSN with data for secondary analysis. It would probably be worthwhile to do some further analysis, since the research design is a sound one, and it is possible that treatment effects could be found. If further analysis is done, it should be noted that there are significant differences among experimental groups in terms of PA versus NPA and application versus recertification. These differences must be taken into account when looking for treatment effects.

2.6 RESULTS AND CONCLUSIONS

Effect on Error Rates

For the reasons discussed above, there is no way to tell at this time whether the intervention had an effect on either type of outcome variable: discrepancies or extent of reporting. If re-analysis were to indicate effects, it would still be very difficult to translate these effects into changes in the actual QC error rate.

Other Benefits

A potential benefit, to both the State and FNS, is the existence of a videotape and a brochure. These products are of adequate quality, and might be used for subsequent demonstrations or corrective actions. Note, however, that there is no reason to believe that the Food Stamp Program in Maryland has any intention to do so.

A second potential benefit could arise from the development of a State database of client characteristics. We do not know whether Maryland has actually developed a database or, if they have, how they plan to use it.

It should be stressed that the demonstration did nothing to upgrade Maryland's research capabilities. All project staff who were involved in research activities were working under temporary contracts and are no longer employed by the Food Stamp Program.

NORTH CAROLINA PROJECT

3.1 OVERVIEW

The North Carolina demonstration was intended to improve the quality of interview data by the use of more structured interview modalities: A Structured Manual Interview (SMI) and a Computer-Assisted Interview (CAI). These products were developed by the Center for Urban Affairs and Community Studies (CUACS) at the University of North Carolina, under contract to the State of North Carolina.

Development of the products, particularly of the CAI, was time-consuming and fraught with problems. More planning at the outset would have been very beneficial. CUACS did a good job, however, of involving state and local staff at all stages of development, and eventually produced a CAI that is at least workable and acceptable to local staff. (The SMI, while "workable" in the strict sense of the word, appears to be neither beneficial nor acceptable, except possibly as a training tool.)

The research design associated with this effort was exemplary; it included random assignment of workers (within three counties) to experimental and control conditions, and use of pre-implementation and post-implementation scores. These scores were based on abstracting case files to develop measures of extent of reporting, of completeness of the application form, and of proportion of documented verifications.

Unfortunately, the outcome scores cannot be directly related to QC error rates. Furthermore, as presented in the states's final report, the actual outcomes cannot even be interpreted in their own terms, since only ratios are given. Nevertheless, it appears that the CAI has a beneficial effect on data quality, at least on the three dimensions that were analyzed, and that this effect is more pronounced for new applicants, as opposed to re-applicants or recertification clients.

CUACS devoted a great amount of effort to studying the amount and distribution of worker time associated with each interview modality. It appears that CAI interview time is not much greater than traditional interview time, and that the CAI leads to less time being spent on verification activities. Overall, the CAI is quite feasible in terms of the Food Stamp Program requirements.

3.2 DESIGN OF THE INTERVENTION

Specification of the Problem

The North Carolina Demonstration was based, not on an analysis of QC error data, but on the assumptions that:

1. Client interviews were lacking in structure, and
2. Increased structure in the interviews would tend to reduce the occurrence of QC errors.

Because of this starting point, the project considered analysis of QC data to be of secondary concern. In particular, they did not pursue the distinction between client error and agency error, assuming that a more structured interaction between interviewer and client would simultaneously reduce both types of error. This hoslitic viewpoint certainly has merit, but it makes it difficult to draw clear linkages from the treatment to the anticipated effects.

The baseline, or "traditional," application form in North Carolina is comprehensive and is formatted in a reasonable manner. Its drawback, or at least its potential drawback, is that it can be used in different ways by

different interviewers and for different clients. Thus, an interviewer can skip certain portions that he believes to be irrelevant, can change the order in which questions are asked, and can record ambiguous or inconclusive entries, e.g., with regard to verification sources.

North Carolina hypothesized that this traditional application form promoted error by permitting interviewers to take inappropriate shortcuts and by failing to allow any quality control of the interview process itself. More specifically, they sought to address the following types of problems:

- Discrepancies in client-reported information between different sections of the application.
- Failure to probe adequately, i.e., to follow up on client responses that are vague or suspect.
- Failure to record client-reported information completely.
- Failure to record verification sources and requirements.
- Arithmetic mistakes.

It should be noted that there might well be other motivating factors, in addition to error reduction, that led North Carolina to focus on structured interviewing in general, and on computer assisted interviewing in particular. For one thing, structure promotes uniformity, which is of some value in itself, and which also serves to promote quality control. A corollary is that computerization facilitates supervising (and evaluating) eligibility workers. In particular, this format records the time spent interviewing, which could be used to provide the State with objective productivity data. Such data were not available and subjective informal reports by counties vary widely. Furthermore, computerized interviewing may have some intrinsic benefits, since it is a first step toward a more automated, and presumably more efficient, processing system. There is no way for us to gauge the relative importance of these different factors, as perceived by North Carolina. It seems clear, however, that error reduction was not the sole point of departure for this project.

Definition of the Intervention

The North Carolina intervention was based on two new interview modalities: a structured manual interview (SMI) and a computer-assisted interview (CAI). As originally conceived, the two modalities were parallel in that they were to incorporate the same structure. The CAI, in other words, was intended to be a literal translation of the SM into computerized form.

The structure was intended to incorporate the following features:

- Skip patterns, i.e., branching logic based on previous responses.
- Probes, also based on previous responses.
- Recognition of discrepancies, provision for backtracking when discrepancies are noted, and the requirement to resolve such discrepancies.
- The requirement to complete all responses that are requested, and, in relevant cases, to provide responses in a fixed format.
- The requirement that the stipulated sequence of questions be followed, with no omissions or changes.

This applies to both the SMI and the CAI. In addition, the CAI was designed to produce various output reports including an interview document to be signed by the client, and lists of follow-up requirements for both the client and the worker. Furthermore, the CAI would perform arithmetic calculations when necessary.

Both modalities were intended for use with both applicants and recertification cases. The project was specifically intended to require workers to administer the same questions in the same way to both types of clients, as opposed to taking shortcuts during recertification interviews.

The CAI was designed to be used with standalone personal computers, and was to be implemented in dBase. Training was obviously an important component of the treatment, and was planned for, with training to take place at both the local offices that participated in the demonstration and the CUACS offices in

Raleigh. Furthermore, there was provision for local offices to review the development of the SMI and the CAI, and to make recommendations to CUACS and to the state. All in all, it appeared that the proposed intervention was a feasible one.

Estimation of the Potential Benefit

The North Carolina project did not attempt to develop a numerical estimate of the treatment effect, let alone translate any such effect to QC terms. Their concern was with an "efficiency" ratio, conceptualized as a benefit divided by a cost, but with both of these factors defined only in relative terms.

The "benefit" was thought of as a relative improvement in observable outcomes, e.g., the proportion of missing data on the applications, the frequency of arithmetic mistakes, etc. The "cost" was considered in terms of two dimensions: time and acceptability. Time was operationalized as a set of "work measurement" variables relating to interview time and processing time, and the work measurement study was a major component of the project. There was no expectation that the interviewing and processing time would decrease with the introduction of the new interview modalities. It was hoped, however, that the increase in time would be minimal, and would be acceptable to workers and supervisors.

More generally, acceptability was an important criterion in its own right. North Carolina was concerned with the attitudes and opinions of workers, supervisors and clients, and planned to conduct pre-implementation and post-implementation attitudinal surveys. These were to address the issue of time, as well as other issues: the quality of the interview, the completeness and accuracy of the information obtained, and the discomfort, if any, caused by computerization.

To summarize, the North Carolina project hoped to demonstrate that:

- The new interview modalities led to more complete and more accurate data.

- The increase in time associated with the new modalities was minor in comparison with the improved quality of the data.
- The new modalities were perceived as feasible and acceptable by those people who would be using them.

Thus, the project was not focused on detecting a reduction in QC errors and showing that the reduction in QC errors was cost effective. Its goal was to develop and implement new interview modalities, and show that these were efficient in the sense of yielding better data for a minor increase in worker time.

3.3 RESEARCH DESIGN

Definition of a Control Group

Since North Carolina and CUACS were not primarily interested in QC error rates per se, they did not make the mistake of designing a research model that relied on comparing past and present QC samples. Instead, they developed a true experimental design that incorporated not only randomized selection, but also pre-implementation and post-implementation scores.

The first point to note in discussing the North Carolina design is that it was implemented in three counties--Wake, Person and Almanace--which, as far as we know, were not selected randomly. (The implications of the county selection are discussed below, in the Generalizability subsection.) Workers, however, were randomly assigned to treatments within each county.

In Wake County, a worker handles either applications only or recertifications only. This permits stratification on case type. The project used this stratification variable, and assigned two applications-only workers to each group-- control, CAI and SMI--and one recertification-only worker to each group.

In the other two counties, each worker handles both applications and recertifications, so that no stratification was possible. Furthermore, these counties are too small to permit multiple assignments to each treatment.

Hence, since the CAI was the treatment of greater importance to North Carolina, workers were assigned to a CAI group and to a control group. In Person County, two workers were selected for each group; in Almanace County, three were selected for each group.

These sample sizes--9 workers in Wake, 4 in Person, and 6 in Almanace--permitted a small number of workers in each county to be held in reserve, for possible use as replacements in the event of worker attrition. It was believed, therefore, that the sample size, although small, could at least be maintained. Note that this issue was a serious one: since pre-implementation data were to be used, it was necessary to have workers who were in place before the demonstration as well as during the demonstration. Also, it seemed desirable to utilize experienced workers rather than workers who were in a training phase.

Within each worker, a target sample size of 260 cases was set: 130 pre-implementation and 130 post-implementation. There was no issue of sampling post-implementation cases, since the project planned to use all cases that were encountered during the demonstration period. For the pre-implementation sample, a fixed starting date was set, and the plan was to use cases that were interviewed from this date forward until 130 had been reached for each worker.

This design is exemplary in terms of validity. However, since the worker is the natural (and conservative) choice as the unit of analysis, the sample size is obviously very small. The analytic ramifications of this are discussed below, in the Sample Sizes subsection. For purposes of the present discussion, it is worth noting that other designs might have been used. It would have been possible, for instance, not to have control workers and CAI workers, but to have each worker use the CAI. Appropriate sequencing, i.e., having some workers serve as controls first and then as CAI workers, with the order reversed for others, would have doubled the sample size. The project did not, as far as we know, consider this possibility.

A second alternative would have been to augment the worker sample and reduce the number of cases per worker. This would have been preferable from a

statistical standpoint, but would have required dealing with more counties and training more workers. The project felt that this would have quickly become unwieldy; they were probably right.

Development of Outcome Measures

North Carolina was primarily interested in demonstrating two effects of the new interviews: more complete client reporting, and more accurate completed application forms. By "accuracy," they generally meant an absence of obvious errors: mathematical errors, key entry errors, inappropriately missing data, unexplained discrepancies, and undocumented verifications. Completeness was taken to mean the reporting of more sources of income, more assets, more liquid resources, and more deductions.

These effects can both be measured by taking data directly from the client application. To do this efficiently, CUACS developed a Client Record Abstraction Form, or CRAF, on which data from the application were to be abstracted. Since CUACS staff were to do the abstraction, there did not seem to be any problems with availability of data: It was only necessary to access worker logs (to find out which cases were interviewed at a given time), and to pull the appropriate client files.

In addition, there were several alternative measures that North Carolina was interested in: one, the use of the standard allowance for utility costs (as opposed to using actual costs) was, like the completeness and accuracy measures, directly obtainable from the application form. A second additional measure was the number of client-reported changes. This was not recorded on the application, but could be obtained from special forms (Form 8590) that were included in the case file. Thus, it seemed feasible to include this measure.

A third additional measure was the proportion of verified items later found to be discrepant. Like the number of changes, this appeared to be recoverable from information in the files. Finally, there was an interest in fraud referrals, and it was decided to measure the proportion of referrals

that were confirmed. (This concern with fraud referrals stemmed from the fact that demonstration project funds were used, in part, to hire additional fraud investigators.)

It should be mentioned that the original intent of the project was to look for any treatment effects that might result from use of the CAI and the SM, and that manifested themselves on the application form. Toward this end, the plan was to abstract a very large number of variables from the application. The project was eventually convinced to concentrate their analysis on a small number of key variables. However, the number of variables that were abstracted onto the CRAF remained large.

In summary, the measurement scheme developed by North Carolina and CUACS was not directly related to the QC error rate. It was, however, relevant to the general goal of error reduction, and was based on a data collection process that appeared to be feasible. Furthermore, it also incorporated a detailed work measurement component which, while not of primary interest from the federal perspective, was consonant with the internal project goals.

Sample Sizes

The North Carolina Project hypothesized that the new interview modalities would have thirteen effects:

- (1) Clients would report more sources of income
- (2) Clients would report more assets
- (3) Clients would report more liquid resources
- (4) Clients would report more deductions
- (5) Clients would use the standard utility allowance less frequently
- (6) There would be less missing data
- (7) There would be more client-reported changes
- (8) There would be fewer mathematical errors
- (9) There would be fewer key entry errors
- (10) There would be fewer unexplained discrepancies
- (11) There would be more documented verifications
- (12) The proportion of verified items found to be discrepant would decrease
- (13) The proportion of fraud referrals that were confirmed would increase

These hypotheses--with the possible exception of the thirteenth--are based on case-specific data, so that the number of cases in the sample becomes a consideration.

Since the SMI worker sample contained only two workers, the hypotheses, although stated for both new interview modalities, were thought of more as applying to the CAI treatment. The comparisons, then, were to be based on eight CAI workers and eight control workers: within each group, there were three workers in Wake County, three in Almanac, and two in Person.

The basic plan was to construct various outcome measures, based on the above hypotheses, for each of the sixteen workers, for each of the two time periods. Thus, let

$X(C, G, W, T)$ = An outcome measure, where

C = County (1 = Wake, 2 = Almanac, 3 = Person)

G = Group (1 = Control, 2 = CAI)

W = Worker (Within Group Within County)

T = Time (1 = Pre, 2 = Post)

Then a change measure for a worker would be given by:

$D(C, G, W) = X(C, G, W, T_2) - X(C, G, W, T_1)$,

and these change measures would be subjected to an analysis of variance, with county and group as factors. (Alternatively, the time-one measures could be used as covariates.) This type of design eliminates variance between workers. It does not, however, reduce the variance due to the interactions between workers and treatments.

Theoretically, there are two ways to consider sample sizes within worker in this context. One is to treat the plan as a two stage cluster design. The other is to treat the within-worker samples in terms of measurement error. In practice, both approaches require similar information: some estimates of the pre-test and post-test distributions of the outcome variables. This information was not forthcoming, and no attempt was made to obtain a scientific estimate of the required sample sizes.

It was generally agreed that the design would not permit reliable detection of treatment effects on any one outcome measure. Even if the

treatment "worked" in a vague way, the room for improvement was too limited. Thus, the proportion of clients who had more income to report, the key entry errors that could be eliminated, etc., were too low for a further decrease to be noted.

It was felt, however, that it would be possible to detect improvements on indicators constructed from several measures: indicators of overall number of items reported and of overall number of errors and discrepancies are the obvious examples. But, again, there was no basis for estimating sample sizes from power analyses, and no attempt was made to do so.

Another strategy that North Carolina proposed was the use of nonparametric tests with the totality of outcome measures. Sign tests, for instance, could be used on the total batch of $16 \times 13 = 208$ change scores. This plan was reasonable, but still provided no basis for setting sample sizes.

The target sample sizes that were eventually chosen--130 cases for worker per time period--were essentially based on a combination of cost considerations and professional judgment. If the design is thought of as a two-stage cluster, 130 is definitely on the large side: it would have been more efficient to increase the number of workers and reduce the number of cases per worker. However, this was not possible. Alternatively, if we think of the number of workers as fixed, 130 is probably adequate to detect a moderate-sized overall effect of the CAI, if such an effect is fairly constant across workers. And, if worker/treatment interactions are large, the worker sample is too small, and no increase in the number of cases per worker could have helped.

Generalizability

As mentioned earlier, the three demonstration counties were not, as far as is known, selected randomly. Therefore, there is no statistical basis for extending any demonstration results to the State of North Carolina, let alone to other states. It should be noted, for the record, that all three counties are in the central portion of the state, and that Wake County is an urban

setting, while the other two counties are rural. There is no evidence that we know of to suggest that these dimensions interact with the effectiveness of the treatments.

More to the point, the successful implementation of the treatments could depend heavily on the attitudes of workers and supervisors. And, because it is only one possible component of the larger-scale computerization effort, the CAI might not be suitable in other state settings. For these two reasons, the generalizability of the demonstration is limited.

3.4 IMPLEMENTATION OF THE TREATMENT

Obtaining of Cooperation

The State of North Carolina appointed a project director who had many years of experience as an eligibility specialist in the state system, and was therefore familiar with local operations and concerns. The director served as a liaison between the state, the local offices and CUACS. She presented the state position on the demonstration to the locals and to CUACS, she relayed local concerns and CUACS progress to the state, and she served as an interface between the locals and CUACS.

The director, by virtue of her defined role and her experience and ability, was able to foster a state of informed cooperation during the developmental phase of the project. (She resigned before actual implementation began, and was not replaced, although her immediate supervisor assumed the role of acting director for the remainder of the project.)

In Almanace and Person counties, the point of local contact was the supervisor. Since the offices in these counties are small, the Eligibility Specialists (ESs) were involved during the early phases of the project, and it was clear that cooperation was not a gesture, but an informed decision on the part of those staff who would be participating. In Wake County, which has a larger office, the point of contact was at a higher administrative level, and

it was less clear that potential participants were making such decisions. In general, however, it was concluded that, at the level of supervisor and above, all of the counties agreed to participate fully.

This should not be interpreted to mean that supervisors and ESs were enthusiastic about the demonstration. They were not. The general view that they expressed was that both the CAI and the SM were likely to be too time-consuming, and it would be very difficult to maintain the current workloads while implementing the new modalities. An allied point was that the more flexible standard interview was quite satisfactory when used by a skilled and experienced interviewer. Underlying these criticisms was a less explicit, but obvious, feeling: The demonstration was threatening, since it undermined the individual control of the interview process.

CUACS did an excellent job of addressing local criticism. They responded to objections, solicited input from local staff, and held "prototyping" sessions at which local staff observed and critiqued the development of the CAI. They were also quite clear on the level of involvement that would be required, and on the specific forms that this involvement would take.

Pilot Testing

CUACS simulated CAI interviews at various points in the development of the CAI. The simulations generally indicated a need for further development, and this led to a sequence of "patches" and other programming modifications.

Some of the difficulties involved the content of particular interview items. These were readily resolved, although the resolutions were apt to cause some awkwardness in the interview flow. Other difficulties were related to probes, skip patterns, and consistency checks: it appeared that the CAI was less sophisticated in these areas than FNS had been led to expect. However, it was hard to define the problems in precise terms, since there were no initial (or subsequent) program specifications. As a result, the internal program logic was, and presumably still is, overly complicated and not readily amenable to modification.

Further problems were caused by the use of dBase III as the programming language. This language is primarily intended for file manipulation rather than the type of coding required for interactive questionnaire development, and is not the best choice for the latter purpose.

Difficulties with the CAI development process eventually led to an assessment by FNS and its contractor, Applied Management Sciences. The conclusions of the assessment were that the CAI was lacking in technical capacity and in user-friendliness, and that the basic problem was that a prototyping approach was taken, instead of a formal requirements analysis being conducted prior to the start of the programming work. Prototyping is clearly an acceptable approach for software development. However, it appeared that the software development staff at CUACS were comprised largely of junior research staff with programming, not systems development skills. Hence, they were too inexperienced to implement a prototyping approach successfully. It was decided at this time that it might be better to remedy the problem by starting from scratch, but that this was not feasible. Hence, CUACS continued to add patches to resolve specific problems.

Project Scheduling

The North Carolina project experienced serious delays. Most of these were related to the development of the CAI. Some delays were purely internal, i.e., were caused by CUACS's difficulties in meeting their own timelines. It should be noted, in this regard, that CUACS programming staff were apparently not working full-time on the demonstration project, at least not on an ongoing basis, but were juggling the time among several projects. This is, of course, a common practice in the contract research environment. However, it did not work well in this particular situation.

Other CAI-related delays were caused by the need to keep modifying the product, because it failed to operate in accordance with plans (e.g., it did not allow for returning to earlier screens to correct errors, and the logic was incomplete) based on recommendations from FNS and its contractor. It is unfortunate that this situation arose, and it would have been far preferable to have spent up-front time on a detailed requirements analysis. It should be

stressed that CUACS did not conduct a "requirements analysis," but it was not conducted from the perspective of developing software, but rather at a more general level. Again, although early evidence suggested that the planning was not comprehensive. FNS and its subcontractor assumed a role of providing incremental guidance and advice, which the North Carolina project could accept or reject. In retrospect, it would have been more desirable for FNS and its subcontractor to be more directive. However, there was a question of the political and practical feasibility of such a stand.

Another source of delay was CUACS's decision to wait for a dBase compiler, CLIPPER, that was not available when expected. Waiting for the compiler was probably the right thing to do. However, since new software is frequently not available when promised, CUACS might have considered an alternative at an earlier stage of the project.

Implementation Per Se

Despite the delays in developing the CAI, its implementation went fairly smoothly. To be sure, there were problems involved in using the final product. These included:

- System failures, which occasionally resulted in lost files
- Difficulties in paging back to previous screens when discrepancies were encountered
- Interview times that sometimes seemed excessive for recertifications and for less complicated applications.

Nevertheless, in overall terms, the computerized interview worked: the necessary data were obtained, and the necessary forms and documents were generated.

Applied Management Sciences conducted a process evaluation during the implementation period, and CUACS administered additional attitudinal surveys to clients, workers and supervisors. The results were generally positive. Clients were satisfied with the new interview modality. Workers and supervisors had mixed reactions, although it should be noted that attitudes

toward computerization improved during implementation. With regard to workers' attitudes, the extreme variability seems more salient than the average satisfaction level. Some workers were not only comfortable with the CAI, but considered it to be enjoyable; others were uncomfortable and hostile.

Variability was also an important factor in the way that the CAI was actually used. Despite the initial plan to equate computerization to uniformity, the final product permitted a reasonable degree of flexibility. Workers were confronted with all questions that were relevant to a particular skip pattern, but could--since alphanumeric fields of any length were accepted as responses to most questions--pass through a section of the questionnaire quickly if they so desired. (There was, of course, no way to force a worker to even ask a particular question, let alone to follow the CAI wording of the question).

There is little doubt that this flexibility is desirable, despite the concomitant loss of uniformity. If the CAI were followed to the letter for all cases, most interviews would be excessively long and would constitute an unpleasant experience for both the worker and the client. It is better to view the CAI as a tool that can be used intelligently by experienced and capable workers.

The structured manual interview, on the other hand, appears to be too lengthy and cumbersome for general use. It was not well received. North Carolina has taken the position that the SM can best serve as a training device for new workers; this is a reasonable framework in which to view the SMI.

3.5 IMPLEMENTATION OF THE RESEARCH DESIGN

Data Collection

Data collection included abstraction from case files and work/time measurement. Abstraction was done for each sampled worker for both the pre-implementation and the post-implementation period. The goal was to abstract 130 cases for each worker for each time. There was a moderate

shortfall for the pre-implementation sample and a major shortfall for the post-implementation sample. Two factors contributed to this. First, because some workers had to be dropped from the sample, it was necessary to go back and do additional abstractions for replacement workers. Second, the number of variables that were abstracted were unnecessarily large, and this slowed down the process. Both of these problems should have been avoided.

In general, the abstraction process went smoothly. Uniform procedures were used, and the resulting data appear to be reliable. However, several variables that were believed to be of prime importance were discarded due to difficulty in abstracting them from case records. These included mathematical errors, key entry errors, unexplained discrepancies, verified items found to be discrepant, and fraud referrals. Other key variables were combined, and the resulting set of analytic variables included only three measures; extent of reporting, number of required items actually recorded, and proportion of documented verifications. The research hypotheses, therefore, were significantly reduced from the original set.

Work/time measurement was based on several approaches, which during development appeared to have validity and reliability problems and were recommended for deletion. North Carolina chose to retain them and most were subsequently discarded after data collection. It appears that simple recording of interview time would have been sufficient, and that the more involved measures were not necessary.

Data Analysis

Data analysis included three major components: effectiveness, efficiency, and effectiveness/efficiency ratio. Effectiveness was operationalized in terms of within-worker ratios of post-scores to pre-scores on the three analytical variables. Ratios were computed for each worker for novice cases (new applicants), for experienced cases (re-applicants and recertifications), and for a standardized mix of the two types of cases. These ratios were then aggregated across workers within each experimental condition to yield an average improvement score for that condition. Finally, the ratios of CAI improvement to control improvement and of SMI improvement to control improvement were taken as relative effectiveness measures of the CAI and the SMI.

This methodology is valid and internally consistent. The various ratios, however, are difficult to interpret. In the absence of raw data, there is no way to gauge the absolute improvement from pre to post or the absolute advantage of the CAI over the traditional interview. Furthermore, the small sample size--recall that the worker is the unit of analysis--can be expected to lead to unstable results, particularly since variation among workers appeared to be pronounced.

Efficiency analysis was based on interview times and on processing (verification) times. The interview times are exact (clock) times in all cases; processing times, however, are estimates, and may not be reliable. At any event, total times (interview plus processing) were estimated for each experimental condition for intake and for recertification interviews. These were used to estimate global times for the life of a case as a function of number of recertifications. This is a valid and appropriate strategy to use.

Effectiveness/efficiency ratios were computed by simply dividing effectiveness measures by efficiency (global time) estimates. These ratios have some possible interpretation in that a large value is better than a small value. They do not, however, have any meaning in the cost/effectiveness realm, since their numerators cannot be translated into savings or improvement terms.

Reporting

The final report submitted by North Carolina has several excellent features: it describes the treatments, the data collection procedures, the analytic steps, and the findings in good detail. However, there are a few serious deficiencies, the most important of which is the omission of any summary of the obtained data. Specifically, the report does not present mean scores by worker--or at any other level--on the three derived measures that serve as outcome variables. Without the mean scores, there is no way to assess the practical significance of the resulting ratios.

Furthermore, it is not completely clear how the three derived measures were constructed. Some more detail is needed on this.

Finally, the report does not seriously address the development of the CAI, the possible shortcomings of this product, or the issue of mechanical adherence to an interview protocol versus professional use of the protocol as an aid to interviewing. These topics, discussed earlier in the present report, should be given serious consideration in any future work of this sort that FNS contemplates funding.

3.6 RESULT AND CONCLUSIONS

Effect on Error Rates

In terms of effectiveness, the relative improvements for the CAI were greater than for the control group. This is difficult to interpret for three reasons. First, no hypothesis testing was done--or, at least, none was reported. Second, the ratios that were reported are themselves impossible to interpret in the absence of more data. Third, even if the CAI leads to significant improvement on the three outcome variables, there is no way to relate these variables to QC error.

Despite these problems, the results are quite uniform across outcome measures and across local offices. It seems likely that the CAI does indeed have some positive effects. Furthermore, to the extent that these effects are of interest, they are distinctly more pronounced for novice clients.

In terms of efficiency, the CAI is not as time-consuming as had been feared. With initial interviews and recertification interviews both taken into account, and with adjustment for a saving in verification time for the CAI, it appears that the CAI is a feasible alternative to the standard interview. It also appears that the CAI is acceptable to most workers and clients.

Other Benefits

As stated at the beginning of this chapter, North Carolina never intended to measure the impact of computerized interviews on QC error rates. Their

goal was to show that this mode of interviewing led to some improvement in data quality, and that it was feasible in terms of worker time and acceptability. They succeeded in showing this.

In a larger sense, the CAI will be beneficial to North Carolina if it serves as a starting point for further computerization. Interview data should be updated as verification activities proceed, and should be used to calculate grant amounts. The CAI provides a basis for these functions.

VERMONT PROJECT

4.1 OVERVIEW

The Vermont demonstration originally intended to reduce agency error and client error through four interventions:

- Supervisory case reviews (SCR)
- Staff training program
- Quality circles (QC)
- Performance objectives.

These interventions were developed through analysis of Vermont's QC data, and as an extension and continuation of previous error reduction efforts. Further, these treatments were viewed as being interrelated, with each serving as a foundation for the next. During the course of the Performance Evaluation and Error Reduction Project (PEER), the training program was designed and partially implemented. However, it was concluded that the training seemed redundant relative to other PEER efforts, and was viewed negatively by the caseworkers. Instead of expending additional resources to perfect the training, PEER management, with FNS concurrence, decided to focus resources on the other interventions. PEER management also chose to exclude performance objectives from the PEER project, but to develop performance objectives as an extension of the SCR at a later date.

The demonstration team developed and implemented a sound research design for the SCR. The Quality Circles did not proceed entirely according to the research design, but they did proceed in accordance with the circle process. That is, the circles did not reach the point of testing and evaluating specific outcomes, but rather continued as a process. In conducting the project, key state staff worked closely with a subcontractor, Policy Studies, Inc., to carry out the demonstration and the evaluation.

Although the project schedule was delayed, it was largely attributed to re-focusing and reshaping the project design in ways that made the project more useful to the State and improved the quality of the effort. In general, the project was carried out in an effective manner, with treatment planning, development and implementation carried out as planned, and with evaluation data collection and analysis proceeding similarly. This is not to say that the PEER project had no problems. Rather, it is to say that the problems that occurred were handled competently.

The research design for the PEER project was bounded by Vermont's small size and the fact that from a statistical perspective the error rate was low. The state was too small to produce a sample large enough to detect statistically significant reductions in error. Hence, while the SCR treatment was found to reduce errors, the reductions could not be shown to be statistically significant. The Quality Circles proceeded to the point of identifying problems and solutions, but the project ended before the solutions were evaluated. Hence, treatment effects were qualitative, and positive outcomes were discovered through personal interviews.

This project served as a further step in Vermont's effort to reduce errors by focusing on the role of the eligibility worker in controlling errors. As a result of the project, Vermont has not only adopted the SCR, but has also proceeded with performance objectives. The state continued the Quality Circles, but did not reach a final conclusion about expanding them throughout the state.

4.2 DESIGN OF THE INTERVENTION

Specification of the Problem

The development of the PEER project was based upon efforts taken by Vermont in the past to reduce errors. Most prior efforts were state-level activities and included the implementation of the ACCESS system, Vermont's automated case processing system. Prior efforts in Vermont could be characterized by the state assuming responsibility for error rate and error reduction. The PEER project moved to the local level, focusing on the responsibility of supervisors and caseworkers for error prevention and detection. At the time Vermont proposed the PEER project in 1983, the allotment error rate was 7.5 percent for the most recent fiscal year (October, 1982 through September, 1983). The following six month review period showed an increase to 10.5 percent. An analysis of the QC data revealed that many of the errors cited were identified through a review of documents in the case file. These errors resulted from mistakes made by eligibility workers in initially certifying a household or in collecting information to verify reported circumstances.

In developing the PEER project, Vermont's rationale was that improvements in the quality of work by eligibility workers and supervisors would reduce errors. Vermont noted that while the QC data were useful in making state-wide statements, it masked differences at the level of local offices. The small samples per office made it impossible for local staff to seek to make improvements because the specific nature - or rate - of local error was unknown. The SCR was designed to be a surrogate for a QC review with the major distinction that it would be limited to a desk review, and hence could only detect eligibility worker error. Inherent in the SCR concept was that sufficient reviews would be conducted to allow corrective action to be taken at the level of the unit and the eligibility worker. The SCR was also to be incorporated in to the ACCESS system, to permit automated case selection and to provide timely results to all levels of management. Use of the ACCESS system would also minimize administrative burden in implementing the SCR.

The Quality Circles were chosen as an intervention because of their documented success in improving performance in the private sector. They were viewed as a mechanism for involving eligibility staff in the development and implementation of error reduction strategies, thereby conveying that their performance directly affects the quality of the Food Stamp Program in Vermont.

The two interventions were related in that the SCR could serve to assess the impact of Circle improvements. However, the Circles did not mature enough to warrant such an assessment during the PEER period of performance.

Definition of the Intervention

The SCR is a type of desk audit conducted for a sample of paid food stamp cases each month. The PEER SCR differs from typical supervisory case review procedures in several important ways:

- SCR is a formal standardized system that defines all areas of investigation. It includes a comprehensive instructional packet for supervisors, and a formal mechanism for recording findings
- SCR replicates QC reviews in that all paid cases are subject to review, not just cases within an action in the month preceding review
- SCR reviews are comprehensive and are designed to search through the case for errors that may have been present in earlier decisions and still affect the current payment
- SCR employs random and systematic sampling procedures
- All workers' cases are subject to review
- Linkage with SCR and QC review findings so that SCR outcomes can be compared with QC outcomes
- Integration with the ACCESS system so that review status and findings are always current, and analysis can be conducted to produce reports locally as well as state-wide on a monthly basis.

The SCR is a facsimile of the desk portion of a QC audit and hence is expected to discover the same errors and problems as the QC desk review, as well as identify other problems that require a client action.

The Quality Circles were adapted directly from the private sector. They

- Are voluntary
- Involve employees in solving work-related problems
- Employ a disciplined, step-by-step approach to problem-solving.

The Quality Circles in Vermont followed the typical structure of private sector circles. A state-level steering committee was established; facilitators were trained formally in the Circle methodology and process; Circle leaders were selected, and members volunteered. The steering committee comprised of upper management at the state level, were responsible for developing objectives, policies and procedures, and providing commitment and support. Two state staff members received training in the Circle process and served as facilitators to the local offices. Their responsibilities included coordination of the local Circles, liaison with the steering committee, and support activities, such as identifying external resources and assuring that recommendations are presented appropriately to management. Five Income Maintenance Supervisors were appointed as Circle Leaders. Their function was to direct the overall activities of each Circle, ensuring satisfactory progress in problem selection and resolution, and reporting Circle activities and progress to the Facilitator. Circle members volunteer to devote their own time to solving the error problem and suggesting improvement.

Estimation of the Potential Benefit

With an overall error rate of 7.5 percent in the previous period, Vermont's hope was to reduce the rate to the five percent tolerance level. However, Vermont recognized that the errors were not of a single source. Hence, the SCR was designed to reduce all types of agency errors:

- Misapplication of policy
- Failure to use information reported by the recipient in changing a grant

- Failure to verify information as required
- Failure to follow up impending changes in applicant's circumstances
- Failure to follow up inconsistent information
- Computational errors.

It was complemented by the Circles which eventually were expected to identify error sources and problems. Vermont did not specify a predetermined level of error reduction that it hoped to achieve. However, the following specific hypotheses were proposed for the SCR:

Regarding direct effects:

- The effect of single, case-specific corrective actions will decay over time as the probability of the reappearance of an error in a corrected case increases.

Regarding indirect effects:

- The statewide agency error rate will decrease as a result of general corrective actions implemented for groups of errors;
- General agency errors that were the target of corrective actions by individual district offices will decrease.

Regarding error rates as measured by the SCR and by quality control:

- There is a high, positive association between error rates as measured by supervisory case reviews and those measured by the quality control review process (i.e., that the SCR is a good predictor of potential errors).

Regarding the administrative costs and benefits of the SCR:

- Benefits will exceed administrative costs. Thus, the benefit/cost ratio will be greater than one.

There were no hypotheses specified for the Quality Circles. It was planned that the specific corrective actions tested by the Circle would be evaluated, and since the circles were to identify problems and develop solutions, the evaluation design could not specify them in advance.

Hence, the SCR was designed to eliminate agency error sources, thereby reducing the QC error rate in Vermont. It was also expected that this would be achieved in a cost-specific fashion.

4.3 RESEARCH DESIGN

The research design is discussed only for the SCR because the Quality Circle Design was to be developed concurrent with the problems and solutions identified by the circles.

Sample

The SCR research design evolved from careful thinking about the state's goals, the practicality of implementing an experimental design paradigm, and the fact that Vermont's size affected the adequacy of sample sizes. It was Vermont's goal to reduce state-wide error through the demonstration. Hence, the PEER staff wanted the SCR implemented in all counties. There were concerns about equity among supervisors by adding the SCR workload requirement to some supervisors, but not others. And, with only 12 counties, there was concern about a sufficient sample for treatment and control groups.

All of these factors, in concert with the recognition of direct and indirect treatment effects, led Vermont to develop a design in which all counties and all supervisors participated. From a position of practicality, supervisors were required to review five cases per worker per month, producing a sample of 670 cases for 134 caseworkers in the first month. Vermont planned to increase the sample size, but retained the five cases per worker level when it became apparent that supervisors were experiencing difficulty in achieving that rate. The sample was comprised of two types of cases--those selected randomly and those selected on the basis of error proneness. Included in the random sample were all cases that underwent QC reviews.

Outcome Measures

Five types of outcome measures were identified:

- SCR results – error rates and costs identified and corrected
- QC reviews – error rates and costs
- Administrative costs
- Staff perceptions and attitudes.

Vermont hypothesized that the SCR would have two types of effects on errors: direct and indirect effects. The direct effects are the specific results of actions taken on individual cases to correct errors detected by the SCR. The indirect effect is the reduction of errors in cases that received SCRs as well as other cases by correcting more general problems identified in the course of a supervisory review. It was expected that receiving feedback from a supervisor would prevent the eligibility worker from making the same mistake again, and it was expected that supervisors would alert all workers to the types of problems that were found, and how to avoid them.

Costs and benefits were computed on the basis of amortized developmental and operational costs relative to savings from error reduction. And supervisors and eligibility workers were asked about their perceptions and attitudes toward the SCR.

The Vermont project's ACCESS enhancement allowed supervisors to record the results of their review, the nature of the error. With each case serving as its own control, ACCESS computed the savings associated with the removal of the error, and also allowed PEER staff to examine decay in error reduction over time by examining the SCR record of errors in cases that were selected more than once.

Analytically, PEER staff made comparisons of the case and dollar over-issuance and under-issuance error rates for cases selected randomly (QC and non-QC) and cases selected on the basis of error proneness. Comparisons were also made of errors detected by the QC review versus the SCR, and cost-benefit ratios were developed.

Generalizability

Since Vermont implemented the SCR state-wide, the findings are generalizable to the State of Vermont. Generalizability to other states is limited because Vermont is largely rural and has no major urban area. No analyses were conducted to determine whether variability in demographics within Vermont were associated with different outcomes. Another factor that limits generalizability is that the SCR is tied intimately to the ACCESS system. Hence, without sophisticated computer support, the SCR could not provide ongoing, rapid feedback to management.

Vermont implemented these treatments in the context of a relatively low error rate and after the state had taken many other actions. Hence, the effects achieved in Vermont is also likely to be related to Vermont's evolution and maturity in developing solutions to the error problem.

While it is improbable that these treatments will produce the same level of error reduction that Vermont achieved, the concepts that Vermont developed and tested are highly transferable to other states, and it is likely that the SCR will aid in reducing agency errors in other states. Vermont's exemplary implementation, including its careful review of the problem, the treatment design and implementation strategies, the State's commitment to the project and involvement of local staff, and collection of appropriate outcome data, make the PEER project process highly desirable for transfer.

4.4 IMPLEMENTATION OF THE INTERVENTIONS

The PEER project was carried out under the leadership of an individual, who during the course of the project was appointed the Commissioner of the Department of Social Welfare (DSW). Other key DSW staff played active roles in the project's design, definition and execution. Vermont's commitment to the effort was characterized by active involvement of DSW's leadership and key staff. The PEER staff also included a subcontractor - Policy Studies, Incorporated (PSI). There has been a long history of successful collaboration between Vermont DSW and PSI. PSI was able to provide sophisticated research

and analytical skills. Together, the Vermont and PSI staff worked effectively to deal with the issues that must be solved as a project moves from conception to design to implementation to completion.

Despite these favorable circumstances, the PEER project also suffered significant delays. The advantage of the active involvement of Vermont staff was accompanied by the disadvantage of other state priorities reducing their availability and active involvement. This, however, is a given whenever state staff are involved.

Probably because of Vermont's size and its prior history of local office involvement, implementation of the demonstration proceeded smoothly. In implementing the SCR, PEER staff conveyed to the supervisors that the SCR was well thought out, had face validity as an error reduction device, and was integrated into ACCESS to reduce burden. Their decision to involve all units also enhanced the sense of importance and fairness. Limiting case reviews to five per month also demonstrated sensitivity to the likely perception that the SCR would be burdensome.

PEER project staff developed training materials and conducted a training program in the procedures for review, how to record findings, and how to use the SCR subsystem in ACCESS.

For the Quality Circles, PEER staff performed intense developmental efforts. They conducted a literature review and interviewed people involved with private and public sector applications. Following this familiarization, two state staff attended a training program on how to facilitate and implement Quality Circles. The Circles were carried out in accordance with standard procedures. Five of Vermont's district offices volunteered as sites, and four were chosen. In general, the membership in the circles was retained.

Overall, Vermont's implementation proceeded smoothly and competently. The PEER staff were effective in gaining necessary cooperation, and were supported with a strong commitment from top State management.

Project Scheduling

Like the other demonstrations, the Vermont project suffered from significant delays. There were two general sources of delay. First, the issues that all researchers face when carrying out a complex project had to be dealt with. In Vermont, additional time was spent on specifying appropriate outcome measures for the direct and indirect effects, deciding whether to implement the SCR in all or some offices, etc. The interventions by FNS and Applied Management Sciences in providing review and feedback also contributed to the delays, because the issues raised had to be addressed by the PEER staff. During the PEER implementation, major FNS regulatory changes required state staff to attend to these matters. This also furthered the delay. Overall, while delays occurred in Vermont, they generally were productive and resulted in enhancing the project's quality.

4.5 IMPLEMENTATION OF THE RESEARCH DESIGN

The research design implementation in Vermont was characterized by careful review, assessment and refinement. As is true in any demonstration and evaluation, initial plans are often modified as realities unfold and as careful scrutiny is given to carrying out an idea.

In Vermont Applied Management Sciences contributed to the research design by developing a design issues paper and by developing an error-prone model to target SCR at cases for review with a greater likelihood of error. These products were considered by the PEER staff and were integrated into their implementation. This included distinguishing conceptually the direct and indirect effects of the SCR and developing appropriate measures and analyses. At FNS' request, PEER staff also placed more emphasis on cost-benefit analysis.

The implementation of the SCR evaluation proceeded smoothly. With its incorporation into ACCESS, outcome measures were easily and routinely extracted. These were augmented with additional statistics routinely generated by ACCESS (e.g., trends in QC error rates), as well as nonstructured interviews with supervisors and caseworkers to obtain perceptual data.

The PEER project was characterized by the research staff's willingness to investigate issues and alter plans as appropriate. An example is that findings showed inconsistency between SCR reviews and QC reviews. PEER staff pulled the specified discrepant cases, examined them, and discovered the source of the discrepancies. This finding was used in a formative manner, to define the SCR procedures more consistently with the QC reviews.

The research design implementation for the Quality Circles was limited to collecting perceptual data from those involved, and documenting status and progress. The PEER staff modified their research plans to collect data that would enable them to make meaningful statements about outcome.

Data analysis proceeded in accordance with the plan for analysis. The PEER staff performed a competent analysis, investigated issues that emerged as preliminary output was viewed, and developed reasonable interpretations of the findings.

Reporting

The Vermont PEER project final report was well prepared. It describes the treatments, their rationale, summarizes project procedures, the analyses and findings. The report is sufficiently detailed and generally documents the methodologies that were used. Its deficiencies lie in its lack of discussion of adequacy of sample sizes and details about the treatment development and implementation.

4.6 RESULTS AND CONCLUSIONS

Evaluation results showed decreases of 20 percent, 24 percent and 23 percent, respectively, in overall agency case error rate, overissuance error rate or total agency allotment error associated with the SCR. Although the three error rates dropped during the demonstration period, these changes were not statistically significant. This was attributed to the very small number of observations on which the statistical tests of means was based, however. Interviews with supervisors and workers suggested, moreover, that the SCRs impact on error rates resulted from both direct and indirect effects.

The evaluation also analyzed the SCRs effect on the reappearance of errors in corrected cases. Results indicated that SCRs would reduce the quality control error rate by 8.7 percent. It was expected that this percentage would increase over time, as workers and supervisors became accustomed to the procedures and corrected errors they found in error prone cases.

The evaluation design postulated that supervisors would use SCR results to formulate broad corrective action plans. The evaluation determined, however, that in most cases it was difficult for supervisors to identify problems that reoccurred over time, on the basis of SCR findings. Errors tended to be too distinct to identify general patterns. Furthermore, supervisors would identify a need for general corrective actions on the basis of one or two errors. When interviewed, supervisors also had a difficult time associating the corrective actions they had taken, such as approaching individual workers or bringing errors to the unit's attention during staff meetings, with the identification of particular errors. This may be explained in part by the fact that SCR reports were available three months after the month of review. Actions may already have been taken in the interim period on the basis of recollection.

Results of the cost benefit analysis of SCRs showed that an SCR was cost effective only if its benefits for both the Food Stamp Program and the State 5 AFDC program are considered. The cost/benefit ratio was 1.83. The benefits accrued to the Federal government, since Food Stamp allotments are 100 percent federally funded. None of the cost/benefit ratios exceeded 1 for the state, since it pays none of the allotment costs and half of the administrative costs. State 5 expected future savings from the SCR system to exceed costs, however. When potential reductions in the sanction level as a result of SCRs were considered, moreover, the benefit/cost ratio exceeded 1 for both the Federal government and the state.

A comparison of SCR findings and QC findings found that the ability of SCRs to detect errors also found by QC reviews improved substantially over the demonstration period. The disparity between SCR and QC findings fell from 7.2

to 2.8 during the demonstration, for cases in which SCRs detected errors also found by QC. This discrepancy was expected to fall even more as supervisors made SCRs a greater priority and improved their review procedures.

Information on outcomes of the Quality Circles was not reported by Vermont because the circles had not completed a solution during the life of the PEER project. Quantifiable benefits were expected, however, once recommendations have been implemented. The estimate for annual operating costs was \$70,014. The process analysis of the Quality Circles indicated that while workers derived a heightened sense of teamwork from the Quality Circles, most felt that participation required too much time, given the other demands of their job. Few workers devoted more than one hour per week to the Circles. Most also felt, however, that this one hour was insufficient to accomplish what was required for effective action. A staff survey also indicated that most felt Quality Circles had not achieved any of their intended objectives, at the time of the evaluation. None of the Quality Circle's projects had been completed at the time of the evaluation, however. One of the key lessons that had been learned from the demonstration of this treatment, was the importance of the facilitator and leader roles in keeping the Circle focused on its goals and maintaining group momentum.

The SCR has been incorporated into the Vermont procedures. The project produced useful products for the state, as well as for other states. However, the linkage of the SCR to ACCESS, limits its direct transportability. The Quality Circles were well documented, and the products associated with them would be useful to other states desiring to consider this approach.

COMPARATIVE ASSESSMENT

Although the three state demonstrations funded by FNS shared the common goal of error reduction they differed in many important ways - in design, in execution and in outcome. The purpose of this chapter is to review the similarities and differences among the three demonstrations, and to relate these to the outcomes.

Origin and Nature of the Intervention

The three state demonstrations differed at onset in how the states chose the treatments. In Vermont the treatment was selected as a continuation of building upon what the state had done and treatment was integrated into the state's operating systems. The North Carolina project was conceived by the subcontractor and accepted by the state. Unlike the Vermont demonstration which formalized and refined an existing supervisory case review practice, the North Carolina demonstration was the computerization of the application process - the replacement of a key operational step. The Maryland demonstration was conceived by the Welfare research director and did not involve a subcontractor in executing the demonstration or evaluation. Like Vermont's intervention, the Maryland intervention did not replace a key operational step, but extended existing practices (i.e., used a more sophisticated media approach to explain eligibility requirements). Overall, it appears that the interventions that created the least turbulence were easier to implement.

The demonstrations also differed in who they primarily affected. The Vermont demonstration entered the system at the level of local office supervisors. Supervisors were most burdened by the intervention. Their

participation affected eligibility workers, and in turn affected clients. The North Carolina's computer aided interviews affected eligibility workers and clients directly. The primary burden of this intervention was placed on the caseworker. In Maryland, the media presentations were placed in waiting rooms and hence affected receptionists and clients. This intervention placed little burden on either, but its repetitiveness in busy offices did make it become obtrusive. In examining the ease of implementation, the intervention that affected more senior state staff (supervisors) went most smoothly, and those that affected clients and eligibility workers directly were more difficult to implement. This may have occurred because the supervisors are more likely to view responsibility for error reduction as a part of their job than caseworkers or receptionists.

A key aspect of the evaluation design was the specification of outcome measures. The three demonstrations differed in outcome measures because these were linked to the nature of the treatments and the research hypotheses. The supervisory case review was primarily a detection mechanism - it identified an error and sought to correct it. Hence, the outcome measure was the primary savings associated with the error correction. There was a strong and direct link between the treatment and the outcome measure, and in fact the outcome measurement was an aspect of the treatment. Detecting and correcting errors gave supervisors immediate and positive impact, and an immediately evident effect. The outcomes in Maryland and North Carolina were not as proximate to the treatment. And, their effect would not be likely to be known to the individuals most directly affected by the treatment. Both the Maryland and North Carolina strategies were designed to prevent errors from occurring, not to correct an existing error. In general, prevention strategies are thought to be highly desirable because they have broader impacts and may achieve greater savings. However, by definition, they are not proximate to impact and hence those most affected by the intervention do not receive positive reinforcement about the effect of the treatment. Not only do the prevention strategies not have built-in incentives for the research subjects, it is more difficult to measure these effects. Not being proximate to outcomes, the prevention strategies are more susceptible to other factors making a contribution to an outcome. Hence, prevention strategies require more rigorous research controls. This, in turn, often translates into more complex research

designs, larger samples, more control groups, and more difficult measurement and data collection. Both the Maryland and North Carolina analyses of outcomes eventually broke down due to problems involving outcome data weaknesses and/or insufficient samples. Vermont, with a more proximate measure and a built-in data collection system (ACCESS) was able to achieve its analytic goals.

Sophistication of State Infrastructure

Vermont can be viewed as a more mature state in its service delivery and management information systems. In the absence of an automated management information system, Maryland had to launch a substantial primary data collection effort. While North Carolina does have automated support, the project's research needs outstripped the system's capacity and North Carolina staff also had to expend substantial effort in data abstraction and collection. Hence, the state's sophistication and maturity also played into the demonstration's success.

Project Organization

The three demonstrations differed significantly in staffing and organization. The Vermont demonstration was directed by a senior state manager and involved several key staff in developing and implementing the treatments. This organization represented a strong commitment to the project. Vermont was supported by a capable subcontractor that assisted where needed, but had primary responsibility for the evaluation. The North Carolina Project Director was also a senior manager. However, the leadership duties were delegated to a more junior individual hired specifically to provide day-to-day management. The North Carolina project did not evidence high level involvement from top management. Other state staff did not play key roles, and the state's subcontractor had primary responsibility for the project. Problems emerged when the state's day-to-day project manager had to oversee a much more senior subcontract director and mediate between FNS and the subcontractor.

In Maryland, the project was located in the research office and was headed by a mid-level manager. No other regular state staff were assigned to the

project. Instead, it was staffed by contract employees. Although the level of leadership in Maryland was adequate, the lack of integration between the research office and the other state offices that was needed to support the project hampered data collection and data quality.

Implications

This discussion suggests that demonstration success is linked to the following features of the state agency:

- A strong involvement, commitment and leadership for senior state staff
- Assignment of state staff to the project
- Integrating the intervention into ongoing systems
- Sophisticated management information capabilities and demonstrated linkages and cooperation from other involved state and local agencies.

This discussion also notes that detection interventions that are integrated into existing systems and affect supervisory office staff are more amenable to successful implementation and evaluation. This is not to say, however, that these types of interventions are most successful at error reduction. The implication is that more effort, more support, more time and a larger budget is probably necessary to carry out a prevention mechanism successfully. This suggests that in considering future demonstrations, FNS pursue one of two avenues:

- Select proposals that show promise of implementation success; or
- Be prepared to provide the resources necessary to take a good raw idea and work it into a quality design and implementation.

Volume III discusses how FNS can structure the demonstrations to assure this success.